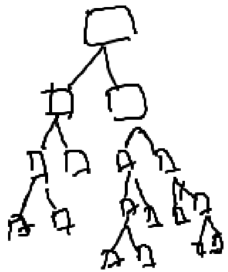
Drzewa decyzyjne

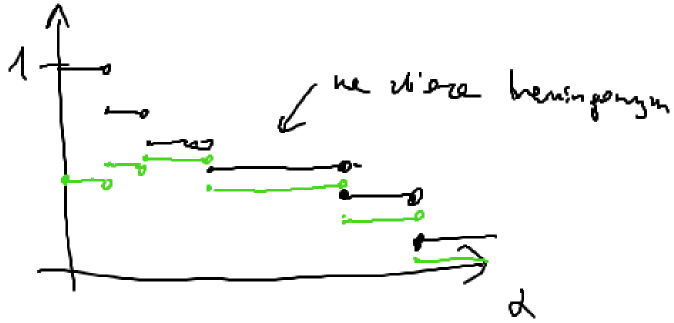


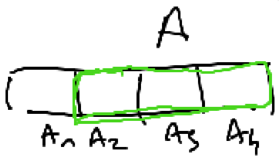
Przycinanie – parametr α

Jak znaleźć α?

1) wykres dokładności na zbiorze testowym

zwykle wygląda tak:





2) walidacja krzyżowa: dzielimy dane treningowe na K części: $A_1, A_2, \dots, A_K$

K-krotnie powtarzamy, dla $k=1,\dots,K$:

ustalmy α_0.

i) tworzymy drzewo dla danych treningowych bez A_k

ii) przycinamy je dla α_0.

iii) patrzymy na dokładności na zbiorze A_k

$\Rightarrow$ wybieramy dobre α_0 i tworzymy drzewo ze wszystkich danych treningowych

• Bootstrap aggregation (bagging)

Ogólna metoda, ale najczęściej używana do drzew.

1) Bierzemy B zbiorów treningowych wielkości N każdy

2) Trenujemy model dla każdego ze zbiorów

3) Łączymy je w jeden model:

   • biorąc średnią dla regresji

   • biorąc „głosowanie" dla klasyfikacji

np. N = 1000 elementów

losujemy 1000 krotnie element ze zwracaniem

ad 1) Każdy ze zbiorów treningowych jest stworzony z oryginalnego zbioru poprzez losowanie próbek ze zwracaniem N krotnie (najczęściej N = wielkości oryginalnego zbioru)

ad 2) W przypadku drzew przycinanie nie jest konieczne, bo i tak później uśredniamy.

N elementów $\{1, 2, \ldots, N\}$

$$X_i = \begin{cases} 1 & \text{, jeśli } i \text{ zostanie wybrane choć raz} \\ 0 & \text{, —||— nie zostanie wybrane ani razu} \end{cases} \quad \text{w } N \text{ losowaniach}$$

$$P(X_i = 0) = \left(1 - \frac{1}{N}\right)^N \qquad EX_i = 1 \cdot P(X_i = 1) + 0 \cdot P(X_i = 0) = 1 - \left(1 - \frac{1}{N}\right)^N$$

$$E(\text{liczba różnych wybranych el. w } N \text{ losowaniach}) = E(X_1 + \ldots + X_N) = E(X_1) + \ldots + E(X_N) =$$

$$= N\Big(1 - \underbrace{\left(1 - \tfrac{1}{N}\right)^N}_{\downarrow \atop \frac{1}{e}}\Big) \underset{\text{dla dużych } N}{\approx} N\underbrace{\left(1 - \tfrac{1}{e}\right)}_{\approx 0.632}$$

Mądrość tłumu: na jarmarku poproszono uczestników o oszacowanie wagi wołu (konkretnego)

średnia ~~[illegible]~~ podawana waga okazała się być zaskakująco bliska prawdziwej

(Bagging trochę to przypomina)

Lasy losowe (?) (random forests)

obserwacje $x_i \in \mathbb{R}^p$

Jak bagging, ale w kroku 2):

$m \in \{1, \dots, p\}$

2') za każdym razem, niezależnie wybieramy losowo pewien podzbiór zbioru $\{1, 2, \dots, p\}$ o m elementach i tworzymy drzewo na danych treningowych jak poprzednio, ale używając tylko niektórych zmiennych, tych o indeksach z wylosowanego podzbioru $\{1, 2, \dots, p\}$.

(wskazówka: $m = \lfloor \sqrt{p} \rfloor$ dla klasyfikacji, $m = \lfloor p/3 \rfloor$ dla regresji ?)

Zadanie 3: - znaleźć dane (np. Breast Cancer)
Titanic

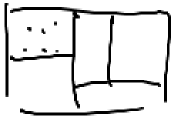
do 9 V

- zobaczyć, jak działają: drzewo, ... z przycinaniem, bagging, random forest
(wszystkich)

Boosting (dla regresji)

N obserwacji próbek $x_1, \dots, x_N$

$y_1, \dots, y_N$



parametry: B – liczba drzew

λ – parametr, dość mały, ale $\lambda B > 1$

d – liczba podziałów w drzewie ($d+1$ = liczba liści)

$$\hat{f}^{(b)}(x) = \sum_{j=1}^{d+1} c_j \mathbb{1}_{R_j}(x)$$

1) Bierzemy $\hat{f}(x) = 0$ i $r_i = y_i$ dla $i = 1, \dots, N$

2) dla $b = 1, 2, \dots, B$:

(a) tworzymy drzewo $\hat{f}^{(b)}$ z $d+1$ liśćmi dopasowując je do danych $(X_i, r_i)_{i=1}^{N}$

(b) uaktualniamy $\hat{f}$ i r:

nowe $\hat{f}(x) = (\text{stare } \hat{f}(x)) + \lambda \cdot \hat{f}^{(b)}(x)$

nowe $r_i = (\text{stare } r_i) - \lambda \hat{f}^{(b)}(x_i)$

$x_1 \rightsquigarrow y_1 = 3.5$ „r_1", $\lambda = 0.1$ $\quad \hat{f}^{(1)}(x_1) = 3.2$

$\hat{f}(x_1) = 0 + 0.1 \cdot 3.2 = 0.32$

nowe $r_1 = 3.5 - 0.32 = 3.18$

$\hat{f}(x) = \lambda \left(\hat{f}^{(1)}(x) + \dots + \hat{f}^{(B)}(x) \right)$

niezmiennik: $\hat{f}(x_i) + r_i = y_i$

Brakujące dane: Możemy:

- pominąć rekordy (obserwacje), w których są brakujące dane
- sztucznie uzupełniać brakujące dane

  |–1–| |–2–|

  1 0.6 K
     0.4 M

  (? , 173,)
  płeć wzrost

- traktować braki jako specjalne kategorie
- dzielić próbki na ułamkowe próbki (C4.5 do drzew)

  • płeć !
  K ↙ ↘ M
  0.6 0.4

  • wzrost < 175?

- można użyć niektórych algorytmów tworzenia drzew decyzyjnych bezpośrednio do danych z brakami (C4.5) (ale nie w scikit)

  Ross Quinlan

Dane niezbalansowane

Np. dwie klasy o bardzo różnej liczności

np. 98% osób zdrowych A −
2% -"- chorych B +

można użyć bootstrap do zwiększenia liczności mniejszej klasy

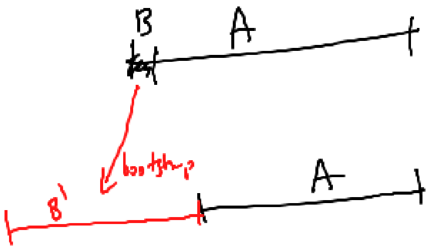


Tabela pomyłek (Confusion matrix)

klasyfikacja binarna

klasa predykowana \ rzeczywista	−	+
−	3125	421
+	35	2235

liczba obserwacji nie chorych

$$\frac{421+35}{421+35+3125+2235}$$